TSS assembly pipeline for Hs_EPDnew_002

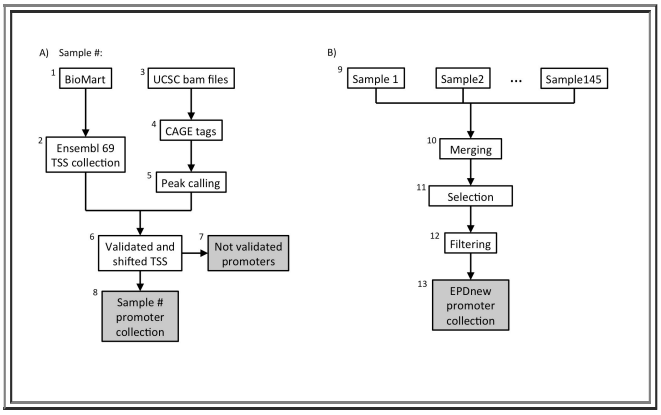
Printer version

Introduction

This document provides a technical description of the transcription start site assembly pipeline that was used to generate EPDnew version 002 for *Homo sapiens* genome assembly hg19.

Source Data

Assembly pipeline overview



Description of procedures and intermediate data files

1. Biomart Download

Data was downloaded from oct2012.archive.ensembl.org/biomart/martview/ selecting the following attributes:

1. Ensembl Gene ID
2. Ensembl Transcript ID
3. Chromosome Name
4. Strand
5. Transcript Start (bp)
6. Transcript End (bp)
7. Gene Start (bp)
8. Gene End (bp)

9. Status (transcript)
10. Status (gene)
11. Associated Gene Name

Then, transcrips have been filtered according to the following rules:

1. Transcripts of protein coding genes only
2. Transcript length > 0 [Transcript Start different from Transcript End]
3. Transcript lies on full chromosomes
4. Genes must be annotated [Associated Gene Name present]
5. Gene and transcripts status known

Gene names were taken from the field "Associated Gene Name". Since the EPD format doesn't allow gene names longer than 18 characters, we checked whether the names repsected this limitation.
Transcripts with the same TSS position were merged under a common ID. As a conseguence of this the total number of TSS in the list was 100276.

2. EMBL TSS collection

The ENSEMBL TSS collection is stored as a tab-deliminated text file conforming to the SGA format under the name:

Hum_ENSEMBL69.sga

The six field contain the following kinds of information:

- NCBI/RefSeq chromosome id
- "TSS"
- position
- strand ("+" or "-")
- "1"
- TranscriptID..GeneName.

Note that the second and forth fields are invariant.

3. Data import from ENCODE CAGE

Solexa Tag Data were downloaded from UCSC ftp-site (see link above). The source

files are in bam format. The complete list of files can be found here. Bam files were converted into bed files with bamToBed program. Files were keps and analysed individually.

4. CAGE tags

The compressed version of these file are available from the MGA archive (see above) under the GEO series ID GSE3444.

5. mRNA 5' tags peak calling

Peak calling for each individual CAGE data file has been carried out using ChIP-Peak on-line tool with the following parameters:

- Window width = 200
- Vicinity range = 200
- Peak refine = N
- Count cutoff = 9999999
- Threshold = 10

6. TSS validation and shifting

Each sample in the collection (mRNA peaks and ENSEMBL TSS) was then processed in a pipeline aiming at validating transcription start sites with mRNA peaks. An ENSEMBL TSS was experimentally confirmed if a CAGE peak lied in a window of 100 bp around it and if it had a maximum high of at least 10 tags. The validated TSS was then shifted to the nearest base with the higher tag density.

7. ENSEMBL not-validated TSS

The total number (summing up all samples) of non experimentally validated TSS was around 75000.

8. Promoter collection for each sample

Each sample in the dataset was used to generate a separate promoter collection. These individual collections were used as input for an additional step in the analysis (Assembly pipeline part B). The aim of this step was to select the promoters that were validated by high number of samples thus increasing their reliability.

9. Merging of the data and second TSS selection

The 145 promoter collections were merge into a unique file and further analysed. The promoter of a transcript was mantained in the list only if validated by at least two samples. This could potentially lead to TSS to be set on a broader region and not to single position. To avoid such inconsistency, for each transcript we selected the position that was validated by the larger number of samples as the true promoter.

10. Filtering

Transcription Start Sites that mapped closed to other TSS that belonged to the same gene (200 bp window) were merged into a unique promoter following the same rule: the promoter that was validated by the higher number of samples was kept.

10. Final EPDnew collection

The 25988 experimentally validated promoter were stored in the EPDnew database that can be downloaded from our ftp site. Scientist are wellcome to use our other tools ChIP-Seq (for correlation analysis) and SSA (for motifs analysis around promoters) to analyse EPDnew database.